Geospatial Data Quality for Analytical Command and Control Applications

Robert F. Richbourg and George E. Lukes
Institute for Defense Analyses
rrichbou@ida.org glukes@ida.org

Abstract

Have you traced a digital representation of a road with so many switchbacks that you questioned the map accuracy? Have you asked an Internet utility to provide a travel route and found the result unintuitive? In each case, flaws in the road network representation may be to blame. Road switchbacks can result from digitization errors such as kinks and kickbacks. Route planning can be defeated by breaks in the network.

Much of the digital map data used to represent the physical environment comes from the National Geospatial-Intelligence Agency (NGA). While the NGA has a large holding of internally-produced geospatial data, the agency's current strategy includes substantial data production under contract and a large cooperative effort with other nations under the Multinational Geospatial Co-production Program (MGCP). The development, codification, and enforcement of detailed quality standards are critical to this acquisition strategy.

This paper uses the modeling and simulation application area to exemplify problems that can arise when digital feature data is used for command and control purposes such as automated route planning. This paper describes the type of quality standards that are to be applied in production of geospatial feature data and illustrates a process to transform semantic descriptions into specific guidance suitable for software implementation. The process includes experimentation to determine appropriate reasoning strategies that will permit identification of substandard data while minimizing false positive notifications. The paper describes the impact on simulation entities using the digital data to exemplify a typical problem, details the experiment designed to address the problem, and presents the results of conducting the experiment. The paper concludes with observations on the potential impact of these geospatial data developments on computer applications that use the data in various reasoning domains.

1. Introduction

The availability of accurate maps has long been a significant factor in the conduct of military operations. "*Know yourself, know the enemy, and the victory will not be at risk. Know the terrain, the natural conditions, and the victory will be total.*" (Sun Tzu, *The Art of War*, circa 300 B.C.) While this tenet has not changed very much since the time of Sun Tzu, the tools available to represent the terrain have changed significantly. Maps have evolved from printed products to electronic raster images to today's digital vector databases. These databases, or vector maps, enable a wide range of capability, including tailored displays and search mechanisms. Some of the more important capabilities are analytical in nature, including intervisability (line-of-sight) predictions and automated route planning. In some sense, these analytic uses are forms of simulation. As such, simulation use of digital terrain data offers a reasonable proxy to explore implications of terrain data quality on operational uses of automated reasoning technologies. This is particularly true for distributed, entity-level simulations that constitute a demanding application area where both syntactic and semantic validity are of fundamental importance. The lessons learned in this simulation domain highlight general issues for geospatial data when applied to analytic command and control problems.

The demands on and expectations for digital representations of the natural environment continue to increase. The current emphasis on urban operations has resulted in a renewed and strengthened interest on many topics that were originally explored when detailed, digital representations of the physical environment first began to appear, such as under the DARPA Synthetic Theater of War (STOW) simulation program started well over ten years ago. Greater numbers of features, represented at higher levels of fidelity and precision, have to be present in the computer representations of the environment if those representations are to be successfully applied to the

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **20 MAY 2008**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE
Geospatial Data Quality for Analytical Command and Control Applications

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses 4850 Mark Center Drive Alexandria, VA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
AFCEA-GMU C4I Center Symposium "Critical Issues In C4I" 20-21 May 2008, George Mason University, Fairfax, Virginia Campus, The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**20**	

areas of interest for today's military operations. Moreover, these features must collectively define a plausible functional representation of the physical world. Today, there are emerging standards for feature relationships that will lead to significantly enhanced digital feature data for use in the construction of digital environments. For example, road features and the bridge features supporting them must be represented as coincident geometries in new National Geospatial-Intelligence Agency (NGA) data.

While such content and feature relationship requirements are difficult to satisfy, there are other challenges. This paper considers examples of both topology and geometric representation, when used in the distributed, entity-level simulation domain, as critical issues where experimental methods can be applied to refine requirements. These areas are particularly relevant because of their potential to ameliorate long-standing issues associated with imperfect feature data in the terrain database generation process.

Topological requirements define the expectations of how geospatial features should be connected, much like the topological requirements for a computer network. As an example, if two individual road line features are intended to comprise a network, then the two line features should have at least one vertex in common so that they connect at that vertex; otherwise, there is a break in the network. Such unintended breaks, based on nearly equal but not identical coordinates at road feature vertices, confound analytical applications such as route planners and lead to computed solutions that are longer than intuitively obvious alternative routes. Interactive (visual) inspection of the digital road network may not be able to detect that a gap exists while a network search algorithm determines that no legitimate route exists between the two vertices.

Geometric representation requirements include expectations for the coordinates of individual feature vertices. For example, individual line features should not have consecutive vertices or segments that are identical (or even nearly identical for that matter). When features include such unintended constructions, unintended consequences usually follow. The simplest consequence is the additional load on application system (graphics or reasoning) when consecutive duplicate vertices must be processed. These may not have any visible manifestation in unusual analytic solutions or in simulated entity behavior, but they do impose an additional computational load that serves no useful purpose. When the duplicate vertices are not consecutive, visibly unusual simulation entity behavior can result. As an example, a simulation entity following a road feature that included certain types of duplicate segments would (typically) execute a movement that seemed to kick back on itself at some point. This might be a nice move on the dance floor, but the "tanker-two-step" is rarely appreciated when the tank is supposed to be executing a movement in a simulation scenario.

Following sections of this paper first describe the emerging requirements that will be applied to NGA geospatial feature data. The paper then discusses line network breaks as an example of a requirement violation that can survive in finished data. Next, the paper describes the motivation for and presents the results of experiments applied to investigate detection mechanisms for a single class of geometric requirement violations. Finally, the paper concludes with observations on the future of simulation terrain database production based on experience with the emerging standards and associated geospatial data inspection utilities.

2. Multinational Geospatial Co-Production Program (MGCP)

Traditionally, NGA and its predecessor organizations focused largely on in-house production of well-defined, predominantly analog, mapping products. This is no longer the case. First, NGA is shifting to the generation, exploitation and dissemination of digital geospatial data, rather than traditional products. Second, NGA has determined that much of their data can originate with other data-producing organizations. NGA no longer depends on exclusive in-house production of geospatial data and out-sources much of its national production.

NGA is also a active participant in the 28-nation Multinational Geospatial Co-production Program (MGCP) [2]. Here, each member nation contributes new geospatial feature data to a common digital data "warehouse." Contributing nations have rights to withdraw data from the contributed pool. Effectively, the member nations have entered into a co-production agreement where each member nation profits from the work of many others. The objective is to complete mapping the majority of the global landmass at scales of 1:50,000 to 1:100,000 by 2011.

The development of standards to rigorously specify both semantic and syntactic quality requirements for the contributed data has been a key concept in making the co-production program viable for all participating nations. The member nations have jointly authored a set of documents to serve in this role. These documents specify both low-level and higher-level requirements for contributed data. Some of the low-level (syntactic) requirements are related to geometric representation and inventory content.

All MGCP data will be represented as two-dimensional point, line, or area geometries that can be included in a file that obeys the ESRI Shapefile [1] conventions. There are many other geometric requirements for each feature. For example, MGCP features may not have duplicate vertices, must obey specific size requirements, must not self-intersect, must not include kink formations, and must be contiguous. The content requirements also state that specific features are mandatory for inclusion in the digital data if they actually exist in the physical world. These types of low-level requirements are highly beneficial to the modeling and simulation community. Some of the semantic requirements offer even greater value.

The semantic requirements for MGCP data are formally described in the MGCP *Semantic Information Model* document that includes such requirements as:

- "An Inland Water Line Feature shall connect to a Water Area Feature by the line feature's terminating point coinciding with the area feature's outer boundary."
- "A Vanishing Point shall be coincident with an endpoint of a River Line Feature."
- "The Dam Point Feature shall be coincident with the endpoint of the upstream and downstream dammed water line feature (river, ditch or canal)."
- "All Road Transportation Feature segments that are connected in reality shall be geometrically connected in the data."

These kinds of requirements directly benefit all communities that would use the data for analytic purposes, such as the simulation and command and control communities. Never before has any similar requirement been levied on the data used to create most digital environmental databases. However, these requirements are merely statements of desired properties unless there is also an inspection capability that can automatically detect violations of the requirements.

NGA committed to the development of such a utility for use by NGA staff, their contractors and MGCP partners. An important aspect of developing the utility has been the translation of the MGCP syntactic and semantic requirements into executable software. That is, human effort has produced the requirements documents to allow comprehension by other humans. The individual requirements are often expressed in inexact terms that convey a general meaning during human conversation, but are not sufficiently specific for direct implementation in an automated inspection system. The remainder of this paper discusses some of the challenges in translating the intent expressed in the requirements documents into a specification that will allow automated inspection and thus guarantee the best possible data in practice.

3. A Semantic Problem: Network Connectivity

One of the semantic requirements exemplified above, "All Road Transportation Feature segments that are connected in reality shall be geometrically connected in the data" is a succinct statement about the representation of road networks in the data set. Clearly, the intent is to ensure captured data that supports network connectivity analyses (including route planning). This simply stated requirement also provides an example that significant disagreement can exist concerning the exact meaning of specific words and phrases. In practice, there are alternate interpretations of "geometrically connected" that differ based on either the exact nature of a "connection" or on the precision used to represent the coordinates of a shared vertex common to the connected features.

.

3.1 Nature of the Connections

It is natural for a human (data consumer) to consider two line features to be connected if they appear to be coincident when the features are displayed (either on a printed map or when drawn on a computer display). A more strict interpretation of "connected" requires that the two features share a common vertex. This more strict definition guarantees that the features will appear coincident during display for a human consumer and also allows automated reasoning systems, not blessed with human visual systems, to readily detect the explicit connection.

The requirement statement does not specify the meaning of "connected" in any way that clarifies the debate. Several data producers interpret the requirement based on the appearance of coincidence, not on the more strict interpretation. In this relaxed view, two line features are connected at points where they intersect regardless of the existence of a common vertex at the point of intersection. The relaxed view leads to several problems for consuming systems.

First, automated reasoning systems not only have to detect all types of intersections between line features when forming network graphs, but have to consider the presence or absence of other features as well. In the case of two-dimensional feature data, road line features that lie underground in tunnels or above ground on bridges have two-dimensional intersections with crossing road line features that lie on the ground surface. A corresponding intersection probably does not exist in the physical, three-dimensional world. Thus, simple intersection detection is not sufficient to determine the existence of a network connection.

A second problem can arise if the data is transformed into a different coordinate system. The NGA data is provided in the geodetic (latitude, longitude) coordinate system. Many applications of the data often require representation in a projected coordinate system such as Universal Transverse Mercator (UTM). Feature data is usually converted from one coordinate system to another by converting the coordinates of the feature vertices (e.g., not by attempting to convert some subset of the infinite number of points between adjacent vertices). In cases where two features are connected based on a terminal vertex of one feature being coincident with a segment (but not a common vertex) of the second feature, it would be unusual for the "connection" between the features to survive a transformation from the geodetic coordinate system to the UTM coordinate system. Thus, preserving network connections, based on the relaxed notion, requires that feature intersection points be detected and added to the features as new vertices prior to transformation. This additional step is rarely taken, but is required to preserve network connectivity.

3.2 Precision of Connecting Vertices

Other problems can arise even when two "connected" features are intended to share a common, connecting vertex. The dilemma here stems from the imprecision in the statement of requirement and the potential for very different perspectives by the data consumer community and the data producer community. Many consumers will interpret the requirement statement to specify that two coordinates can only be connected if they are identical, regardless of the number of decimals used to represent the coordinate. To some degree, this position is based on the desire to support automated reasoning techniques that test for exact equality as part of their algorithms. Basic route planning algorithms may not consider "nearly equal" coordinates to be equal when reasoning about a network of connected features. Thus, it is likely that some algorithms will provide routing solutions that run counter to human intuition when presented with network data that includes very small (imperceptible to the human eye) breaks. Essentially, this view of the requirement states that the data should be made so that it will suffice in all expected-case applications.

The data producer community can view the issue from a very different perspective. Here, a convenient translation of the requirement statement is that two coordinates are equal if they are equivalent internally to the Geographic Information System (GIS) used during the data production process. MGCP has not expressed any requirements concerning any specific GIS to be used in the production processes. The only related requirement is that data be delivered in ESRI Shapefile format. It appears that there are some limitations on GIS capabilities related to data that is created in an internal format and then exported into another (e.g., shapefile) format. It is neither possible to precisely control the number of decimal digits that are written to file nor is it possible to zero fill a coordinate value for all decimal positions beyond a preset threshold. Despite their best intentions, the data producers cannot always guarantee that two coordinates that appear identical in the GIS environment will indeed remain identical after the GIS has finished exporting the data to the MGCP mandated delivery format. "The spirit is willing but the body is weak."

Some attempts to resolve this consumer – producer dilemma rest on snapping all coordinates to a predefined grid where the grid spacing is equivalent to the desired accuracy of the data (approximately 0.1 meters in the MGCP case). That is, for every feature vertex, find the nearest grid point and make that the new vertex coordinate. Intuitively, this translates all coordinates into a finite set of points and is intended to move points that are very close together to the same coordinate.

This solution strategy is used very widely as it is

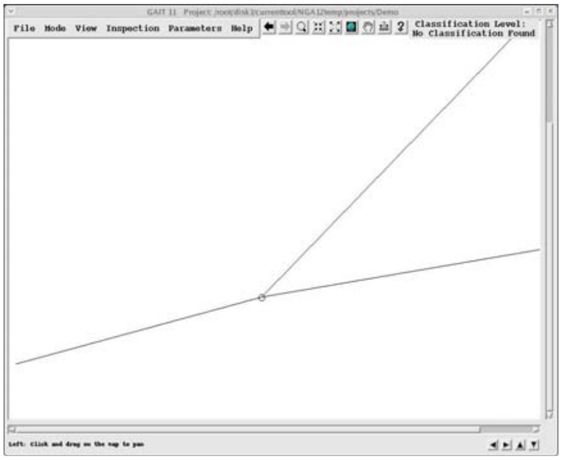
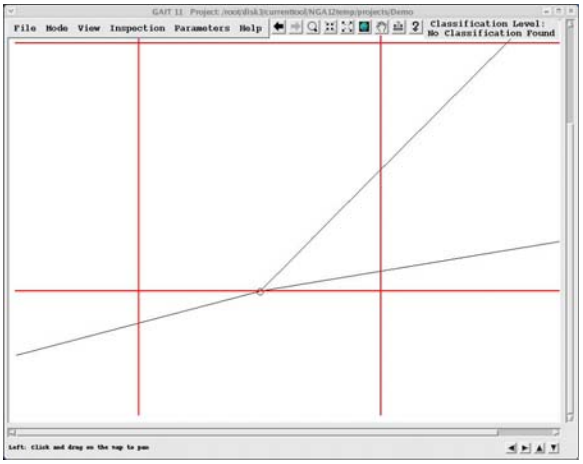

GAIT 11 Project: /root/disk1/currenttool/NGA12/temp/projects/Demo
File
Mode
View
Inspection
Parameters
Help
?
Classification Level:
No Classification Found
Left: click and drag on the map to pan

Figure 1. Road features that appear connected

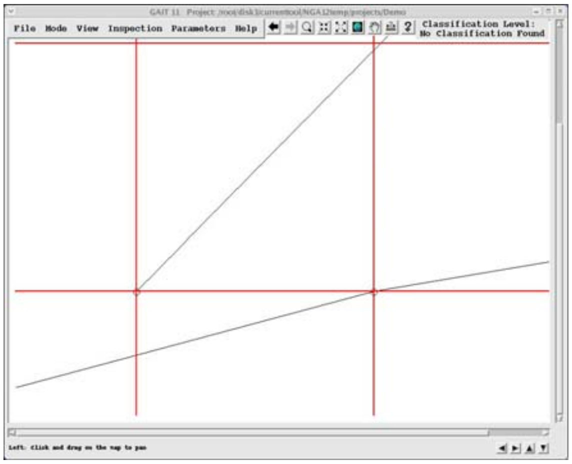


GAIT 11 Project: /root/disk1/currenttool/NGA12temp/projects/Demo

Classification Level:
No Classification Found

Left: click and drag on the map to pan

Figure 2. Adding a 0.000001 degree grid



GAIT 11 Project: /root/disk1/currenttool/NGA12temp/projects/Demo

Classification Level:
No Classification Found

Left: click and drag on the map to pan

Figure 3. Result of snapping to the grid

intuitively appealing and can be applied automatically, as the data is created. However, this strategy also exhibits a characteristic first described by H. L. Mencken: “For every complex problem, there is a solution that is clear, simple, and wrong.” Figures 1 through 3 illustrate Mencken’s wisdom.

Figure 1 is a very tight zoom-in view showing two road features (black lines) that appear to be connected at a common vertex (that has been circled for clarity). The system used to display the features is at the limit of its zoom capability (about 0.2 meters across the entire screen). The two features have a very small difference in their longitude coordinate. The fractional part of the longitude coordinate at the circled vertex on the road coming down from the top is 0.866307500000005.

The fractional part of the longitude coordinate at the circled coordinate on the road moving across the figure is 0.866307499999998. Figure 2 adds a (red) grid to the display; grid spacing is set at 0.000001 degrees (approximately equivalent to the desired accuracy of MGCP data, about 0.1 meters). Figure 3 illustrates the result of having the display system snap the coordinates to the nearest grid point, which in this case makes the difference between the two coordinates far greater, as well as expanding the gap of the network break in the data.

The problem with the approach is that an invisible threshold exists in the grid and this threshold lies halfway between neighboring grid points. Coordinates greater than the threshold will be moved to the next higher grid point while those lower than the threshold will be moved to the closest lower grid point. There will be cases, as shown in this example, where two coordinates that are very close to each other lie on opposite sides of the threshold. Such points will be snapped away from each other, becoming more different from each other rather than becoming identical. Moreover, the finer the grid (less distance between grid points), the greater the likelihood that nearby points will be snapped away from each other because there will be an increased number of thresholds. A finer grid will reduce the maximum distance between “nearby” points, but will not resolve the problem of making those points identical.

A solution to this problem is to examine all of the feature coordinates to select those that are closer together than a set threshold. Once selected, this set of nearby points can be made identical to each other (e.g., not to a predefined grid). Other solutions may exist, but this is the only one shown to work in practice.

3.3 Section Summary

The above discussion illustrates the kinds of issues that must be resolved by automated systems that attempt to enforce the semantic rules designated for MGCP (and thus NGA) data. These same issues will impact systems intended to create environmental representations for use in analytical applications, including simulations. In part, the problems stem from imprecision in the requirements statement; once the standard includes rigorous specifications for the meaning of terms such as “connected”, many of the current problems would be resolved. The MGCP standard is a developing specification, so there is opportunity to apply greater rigor to the semantic rules in the future. In the current situation, there may not always be a “clear, simple” solution, but solutions do exist.

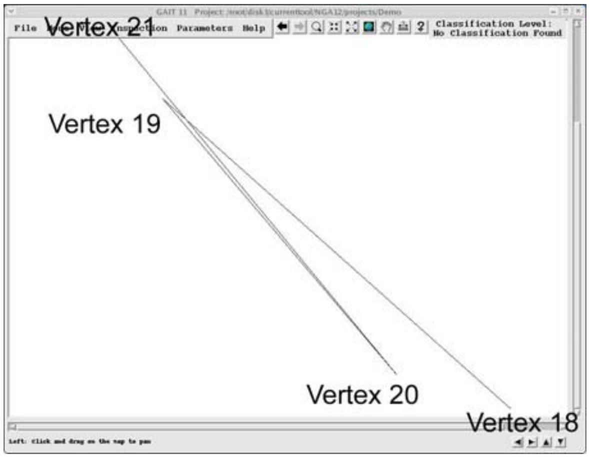


Figure 4. A kink along a road line feature

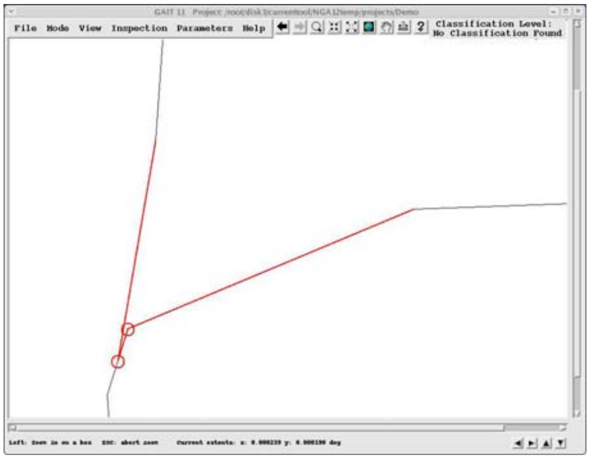


Figure 5. A kink at connected road features

4. A Syntactic Problem: the Line Kink

As above, there are many syntactic requirements placed on the MGCP data. One that is easy to illustrate is that line features are prohibited from having unintentional "kinks" along their length (or at points where two line features connect to each other). A kink is usually caused when a feature vertex is inadvertently entered during the original data capture process. Figure 4 illustrates a typical kink condition that appears to be the result of accidentally entering a feature coordinate (either vertex 19 or vertex 20). Vertices 19 and 20 are approximately 25 meters apart. Vertex 18 is off the screen at the bottom and vertex 21 is off the screen at the top. Imagine the visibly illogical behavior of a simulation entity performing a road-following task along this feature!

Clearly, the geometry illustrated in Figure 4 is not an accurate depiction of a real world road feature. There is universal agreement between data producers and consumers that such constructions should not appear in the finished data. The problem is to devise a procedure to automatically identify such erroneous constructions while minimizing false positive reports of potential errors.

There are alternate methods of detecting line kink conditions [3], but a proven accurate method rests on calculating angles between adjacent segments along line features and at the intersection of multiple line features. Clearly, there are some extreme angles along the length of the feature illustrated in Figure 4 and it is trivial for angle-based detection methods to locate such constructions. The problem with the strategy is that the detected angles are often accurate depictions of the physical world.

Anomaly detectors that have high false positive report rates will tend not to be used. Thus, detection of line kink conditions based solely on angle measurements is not a practical solution. Figure 5 illustrates an extreme angle measured at the point where two road features connect. It appears to be an anomalous construction that should be identified and corrected. Figure 6 illustrates a case where an even more extreme angle has been measured at the connection between two railroad features. Despite the more extreme angle, this appears to be a rational depiction of physical circumstance that would be viewed as a false positive if reported. The following sections describe experiments that have been conducted to help determine a proper mix of angle-measurement-based strategies for line kink identification.

4.1. Refining the Basic Strategy

A comparison of Figures 5 and 6 immediately points to at least two differences between the kinds of line kink conditions that should be reported (true positives) and those that should be ignored (false positives). First, some feature types will always have extreme angles where they connect to other features of the same type. Railroads always have relatively extreme angles at the connecting vertex shared by two features; vehicles that move on rails cannot perform any turning maneuver except high-radius, shallow-angle turns. Conversely, vehicles that travel on roads make these kinds of (shallow-angle) turns as they merge into new travel lanes, but these maneuvers are relatively rare, as compared to the total number of abrupt heading changes along roadways (turns at intersections).

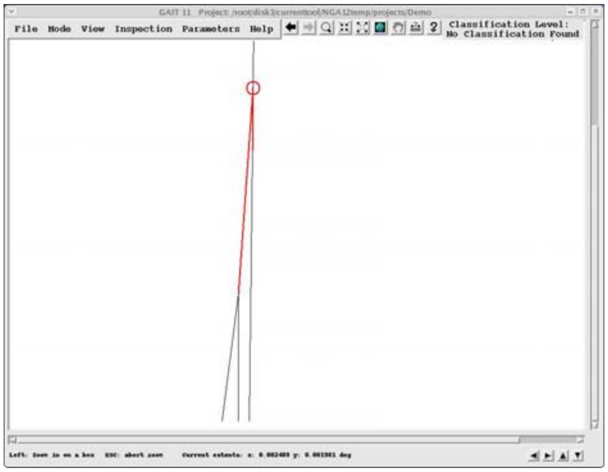
Another readily evident difference between the situations in Figures 5 and 6 can be characterized by the heading of a vehicle moving along the features, immediately prior to encountering the extreme angle at the feature intersection. A vehicle moving along the rail features (Figure 6) would have very small heading changes prior to encountering the feature intersection point. Conversely, a vehicle moving along the road features (Figure 5) would have a significant heading change just prior to encountering the extreme angle at the feature intersection. In practice, both feature type and heading change prior to intersection are viable heuristics in the search for true positive kink conditions.

Experience has shown that two other characteristics can be applied to the problem. First, kink conditions that occur internally to a single feature (e.g., excluding kinks that occur at the point where two different features connect) rarely result in false positive notifications. This heuristic fundamentally depends on a conventional style of representing geometric data. In Figure 6, a conventional style would utilize one (primary) rail feature extending from below the screen area to somewhere above the screen area. Secondary features would then enter the screen area from the bottom and terminate coincident with an interior vertex of the primary feature. An alternate style would represent the geometries as two features shaped like upside-down letter ‘V’ characters. Another (single line) feature would come down from the top of the screen and terminate at the point of the upside-down ‘V.’ The single feature heuristic will not work effectively when faced with this alternate style of representation.

A final characteristic that usually indicates true positive kink conditions also relies on measuring two consecutive angles. In this case, consecutive extreme angles are similar to those that occur in the letter ‘Z’; experience indicates that ‘Z’ shaped features (Figure 7) rarely occur in nature and generally indicate geometric constructions that require some sort of correction. One exception to this rule occurs when transportation features include switchbacks in areas of high slope. Note that this heuristic is very similar to one described above (which considers heading changes). One difference is that the Z-shaped heuristic can be used on single features as well as at connecting points between features (the heading-change procedure applies only to pairs of connected features, at the point of connection).

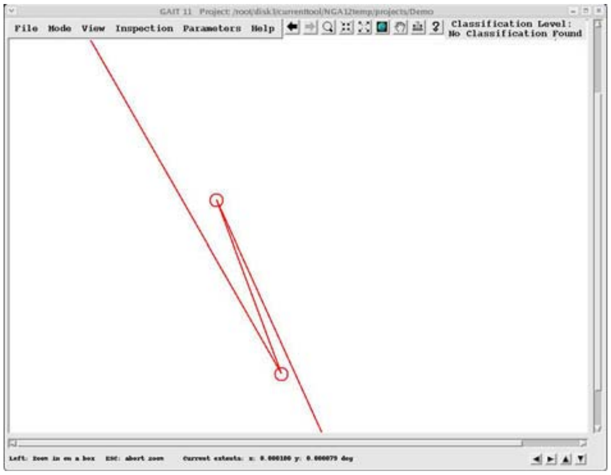
4.2. An Experiment: Judging the Refinements

Original implementations of procedures to automatically detect line kink conditions were based solely on angular measures. This initial, angle-only implementation reported a potential line kink error whenever two adjacent line segments (on a single feature or across a shared vertex between two features) included an interior angle of 10 degrees or less. This procedure was able to identify many true positive kink conditions, but also suffered from two primary drawbacks. First, the true positives amounted to only about 50% of the total reports. Second, many true positives were not detected because they involved kink conditions with angles above the 10 degree threshold. Initial experiments showed that increasing the threshold would result in an increased detection of valid conditions, but also dramatically decreased the true positive report rate, as illustrated in Table 1.



Left: Zoom in on a box ESC: abort zoom Current extents: x: 0.002403 y: 0.001981 deg

Figure 6. A kink at connected rail features



Left: Zoom in on a box ESC: abort zoom Current extents: x: 0.000100 y: 0.000079 deg

Figure 7. A Z-shaped road feature kink

Table 1. Performance of angle-only line kink detection

Thresholds	True Positive	False Positive	Threshold True Positive Rate	Cumulative True Positive Rate
0° < Angle <= 5°	701	236	74.81%	74.81%
5° < Angle <= 10°	676	1033	39.56%	52.04%
10° < Angle <= 15°	628	1929	24.56%	38.54%
15° < Angle <= 20°	744	3018	19.78%	30.66%

The Table 1 results were derived from applying the angle-only kink detector to 14 different data sets. The data sets were chosen to represent a wide range of data that a simulation environment developer might receive as source data from the national holdings. Some of the data sets represented entire countries while others represented high-density urban areas. Together, they included flat desert areas, mountainous areas of high slope, and areas of moderate terrain. In total, the data sets included 1,399,785 line features defined by 51,625,949 vertices. The vertex count is more relevant than the feature count as line kink conditions are always defined by three consecutive vertices (which may or may not involve multiple features). After applying the inspection to each data set to detect potential line kink conditions, each report of a potential error was interactively (manually) reviewed to assess its validity. Thus, the data in Table 1 was created after applying the angle-only line kink detector to 14 data sets and manually reviewing each of the 8,965 reports of potential errors to categorize the 2,749 true positive reports and the 6,216 false positive reports.

Table 1 reveals the unsuitability of applying the angle-only procedure to detect line kink conditions. While the method is capable of locating these conditions, it is not able to support efficient review. A high false positive rate will discourage use of the capability, and thus many of the conditions may not be repaired in the finished data. Also, when the threshold is decreased (from 20 degrees to 10 degrees) to provide a lower false positive rate, the method does not detect over half of the true positives that exist in the data.

Further experiments were conducted to help identify the best mix of heuristics and procedures that would both increase the number of true positive reports and decrease the false positive rate. One of the first experiments explored use of the “Z-shaped” criterion, introduced above, as an alternative to the angle-only procedure. The results of applying the Z-shaped procedure alone are presented in Table 2.

The “Z-shaped” criterion examines three consecutive line segments and the two consecutive angles they form. When both angles are within the defined threshold range, a potential condition is reported. Table 2 shows a very promising performance. The added procedure of considering two consecutive angles allows a higher threshold and reduces the false positive rate; however, this procedure cannot replace the angle-only detector because it is too specialized. It can be used in conjunction with the original procedure. In this case, there is an added concern to ensure that the two procedures do not both report the same condition. Multiple reports of the same problem can be easily suppressed by using mutually exclusive thresholds, as illustrated below.

The two other detection procedures introduced

Table 2. Performance of Z-shaped line kink detection

Thresholds	True Positive	False Positive	Threshold True Positive Rate	Cumulative True Positive Rate
0° < Angle <= 5°	18	1	94.74%	94.74%
5° < Angle <= 10°	127	2	98.45%	97.97%
10° < Angle <= 15°	106	8	92.98%	95.80%
15° < Angle <= 20°	82	13	86.32%	93.28%
20° < Angle <= 25°	95	32	74.80%	88.43%
25° < Angle <= 30°	68	44	60.71%	83.22%
30° < Angle <= 35°	66	41	61.68%	79.94%
35° < Angle <= 40°	39	41	48.75%	76.76%
40° < Angle <= 45°	25	16	60.98%	75.97%

Table 3. Performance of all procedures applied simultaneously

Procedure	Thresholds	Number of True Positive Report	Number of False Positive Reports	True Positive Report Rate
Angle-Only	0° < Angle <= 2°	350	18	95.11%
Z-Shaped	20° < Angle <= 45°	293	174	62.74%
Single Feature	2° < Angle <= 15°	701	320	68.66%
Heading Change	2° < Angle <= 20°	667	249	72.82%

above can be incorporated into the process. These are both variations of the original (angle-only) line kink detection mechanism, so they should exhibit a higher number of detections than the very specialized "Z-shaped" procedure. Again, a concern is that the all of the procedures must be configured so that they do not produce multiple reports of the same problem.

Further experiments revealed that adjustments to the thresholds for each detection procedure can be used to both tune reporting performance and to ensure exclusivity of the reported conditions. Table 3 reports the final results of the entire experiment, using four different detection mechanisms in concert, each with unique thresholds, to identify the line kink conditions.

4.3. Section Summary

The results presented in Table 3 are promising. Original implementation of the angle-only line kink detection procedure used an angular threshold of 10 degrees to identify 1,377 true positive errors out of a total of 2,646 potential error reports (1,269 false positive reports). The refined, combined procedures listed in Table 3 identified 2,011 true positives from a total of 2,772 potential error reports, yielding an overall true positive rate of 73%. Thus, the experimentally refined procedures increased the true positives by nearly half (a 46% increase) while maintaining the total number of reports to be examined at very nearly the same level (total potential errors reported increased by only 4.8%).

Table 4. Procedure performance range

Procedure	Worst True Positive Rate	Best True Positive Rate
Angle-Alone	80%	100%
Z-Shaped	12%	100%
Single Feature	26%	100%
Heading Change	50%	96%

The various procedures showed a range of performance across the different data sets. Table 4 indicates the worst and best performance, by data set, for each procedure. The poor performances of the Z-shaped procedure resulted from its use in two data sets that included many transportation features in steep slope areas. The switchbacks along the features in these two data sets accounted for 87% of all the false positive reports. The Z-shaped procedure performed at a 90% true positive rate across the remaining 12 data sets. Further refinements can be applied to reduce the switchback-induced false positive reports.

The single feature procedure also performed relatively poorly on one data set. Here, the root cause can be traced to the use of atypical feature representations (as described above in ”Refining the Basic Strategy”). The single-feature procedure provided a 76% true positive reporting rate when applied to the other 13 data sets.

There are two experimental limitations to consider. First, although the number of features, and actual errors they generate, are sufficiently large, the number of experimental data sets is low. The problem is that generation of the data sets is somewhat dependent on the style of the data producer (much of the data capture process is manual in nature). Thus, the range of data-creation styles is also low and experiments where other styles have been used may contradict these results. Second, a computer scientist performed the assessment of potential error reports as either true or false positive. A trained cartographer may make other assessments in some cases.

Despite these limitations, the experiment served a useful purpose. The automated procedures to detect a troublesome class of geometric errors on line features have been greatly improved. Other improvements are possible. The procedures could be refined according to the feature type (e.g., road, river, cart track, rail) involved. The experiments described above also captured performance data according to feature type. The analysis of the feature type performance data has yet to be completed. Also, the data shows that the combined procedures did not identify some of the known errors, so there is room for improvement in an overall performance sense (Table 1 shows that at least 2,749 true positives exist while Table 3 indicates that 2,011 true positives were identified). However, the thresholds as reported in Table 3 were selected to find a functional compromise between true positive and false positive reports.

5. Summary and Conclusions

This paper has touched on several separate, but related, subjects. First, there is an increasing demand for digital representations to provide highly accurate models of the physical environment. This is a difficult task, but increased optimism is justified; major data providers are committed to improved geospatial data quality for analytical purposes. That commitment is clearly evident in the efforts of a consortium of nations who have collaborated to produce data quality specifications (both syntactic and semantic) that are without precedent. This is an extraordinary development that will eventually improve the quality of geospatial feature data for the large community of data consumers, including those who develop the modeling and simulation environments. While the specification is still evolving, the potential impact is readily obvious.

This paper has focused on one example of imprecision in the emerging semantic quality specification and traced the ramifications of that imprecision from the context of the data producer to that of the data consumer. There are several other cases of similar concern. In general, the semantic quality specification has been written for human consumption and is not sufficiently precise for direct computer implementation in many cases. The analysis of the single example developed in this paper can be applied to help resolve similar issues of language imprecision and their eventual impact on digital data. The main points are:

- emergence of comprehensive semantic quality specifications is a profound development that will directly benefit analytical users of geospatial feature data;
- imprecision in that specification is problematic; and
- informed consumers need to critically examine data specifications to locate those imprecise descriptions that will impact their use of the data and provide feedback to the production community.

Portions of the related discussion also illustrated the fallibility of the “snap to grid” technique when intended to enforce exact equality of very similar coordinates. Although enjoying widespread use, this technique can only minimize the maximum differences between nearly equally coordinates.

Syntactic issues are also discussed. The paper examines one example to illustrate how structured experimentation (observe, hypothesize, experiment, refine, validate) can be applied to help resolve a problematic issue. The example illustrates that, even though the syntactic requirement is precisely stated and violations of the requirement can be identified, practical issues of supporting efficient review and corrections of the violations must be considered. The main point here is that, despite clear and succinct requirements specifications, there is some probability that prohibited constructions will exist in the finished geospatial data. Practical issues in enforcing data quality drive this circumstance. Again, the data consumer needs to be aware of the limitations of the quality specification. The best specification is of little value without enforcement.

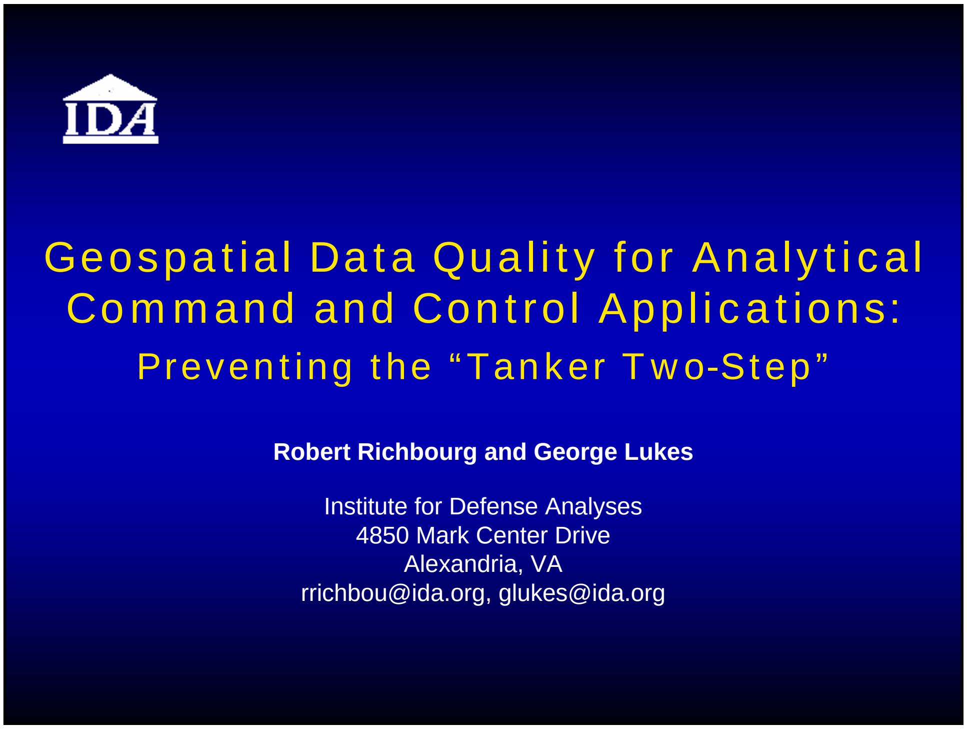
Despite these limitations, there is a solid basis for optimism. Never before has the consumer community seen a comparable emphasis on data quality. NGA is participating in a program to define rigorous and comprehensive quality specifications. Moreover, NGA is leading the effort to implement utilities that can detect violations of those specifications and make routine adherence to the specifications a reality. In the future, the analytic community will include many beneficiaries of this ambitious and timely commitment.

6. References

[1] ESRI, (1998). “ESRI Shapefile Technical Description: An ESRI White Paper”, see http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf.

[2] Fillmore, Randolph (2006), “The MGCP is making big strides towards getting global high-resolution data common across the board”, Military Geospatial Technology, 23 March 2006, V 4:1. http://www.military-geospatial-technology.com/article.cfm?DocID=1380.

[3] Richbourg, R., Lukes, G., and Stone, T., (2004), “Digital Environment Data: Identifying Anomalies from Source to Final Databases”, Paper 1675, Proceedings of I/ITSEC 2004, Orlando, Florida, December 2002.

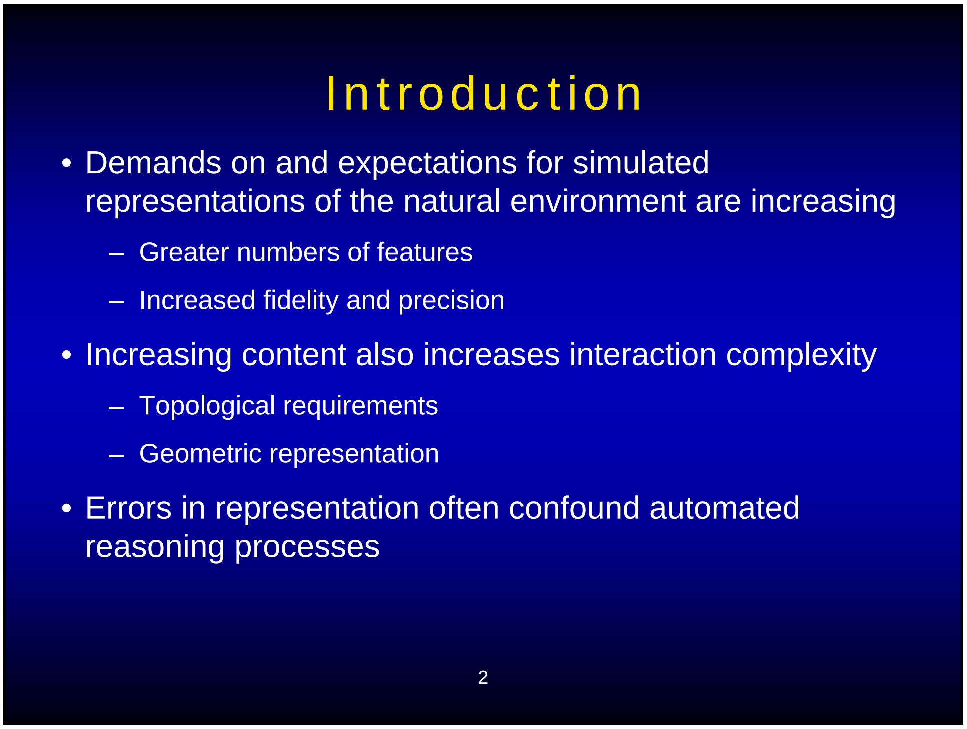


Geospatial Data Quality for Analytical Command and Control Applications:

Preventing the “Tanker Two-Step”

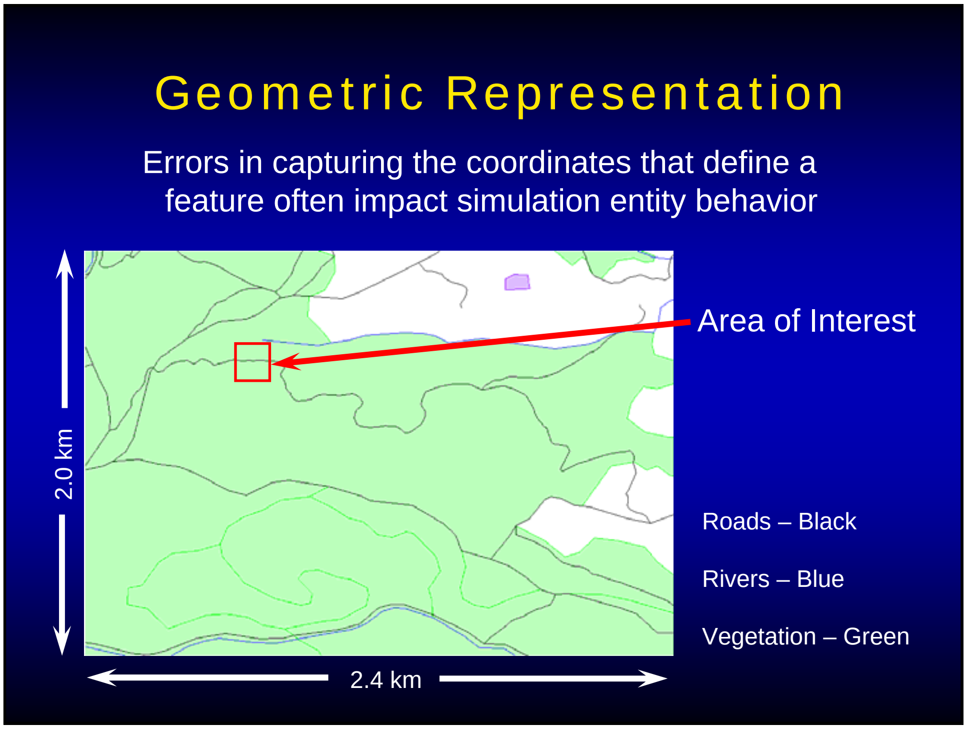
Robert Richbourg and George Lukes

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA
rrichbou@ida.org, glukes@ida.org



Introduction

- Demands on and expectations for simulated representations of the natural environment are increasing
  - Greater numbers of features
  - Increased fidelity and precision
- Increasing content also increases interaction complexity
  - Topological requirements
  - Geometric representation
- Errors in representation often confound automated reasoning processes



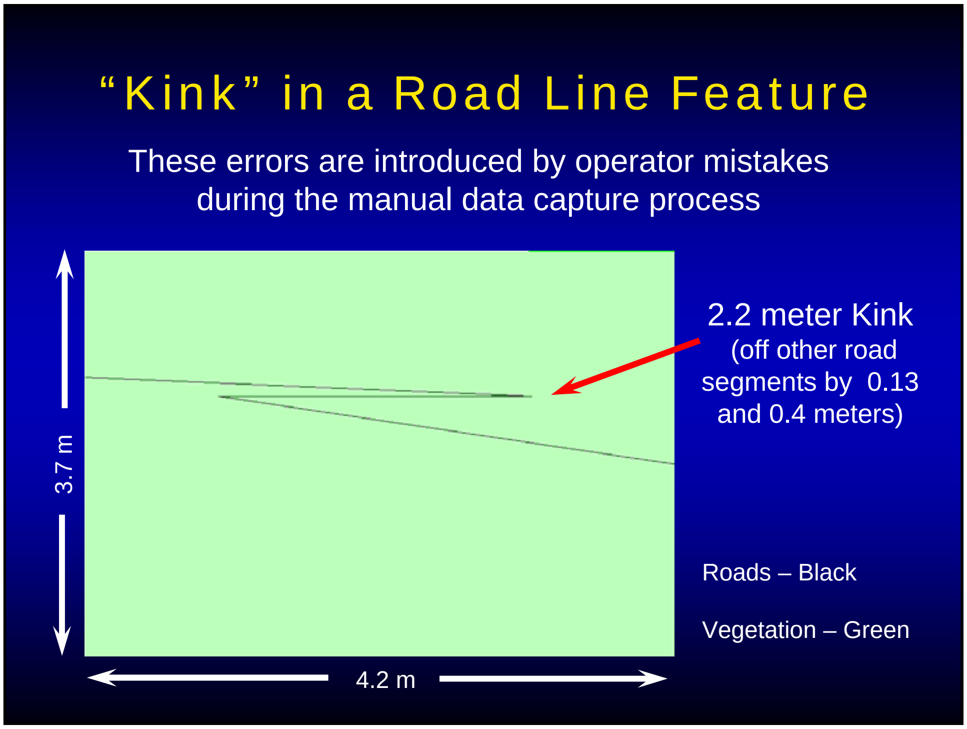
Geometric Representation

Errors in capturing the coordinates that define a feature often impact simulation entity behavior

Roads – Black

Rivers – Blue

Vegetation – Green



“Kink” in a Road Line Feature

These errors are introduced by operator mistakes during the manual data capture process

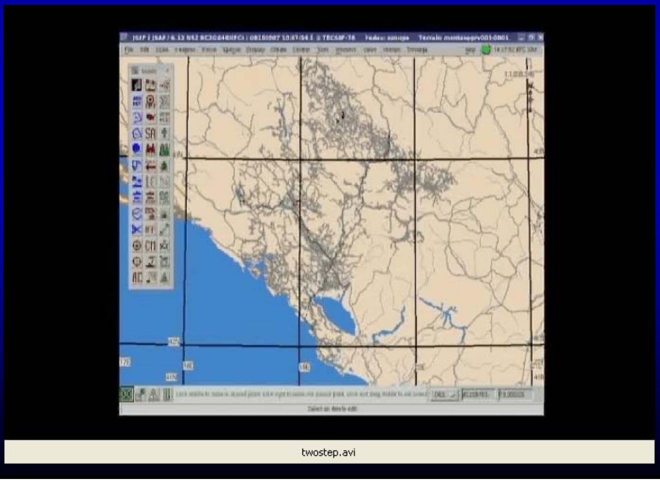
2.2 meter Kink
(off other road segments by 0.13 and 0.4 meters)

Roads – Black

Vegetation – Green

JSAF Tank Entity Behavior

JSAF operator commands a tank entity to follow the road feature across the “Kink” construction …



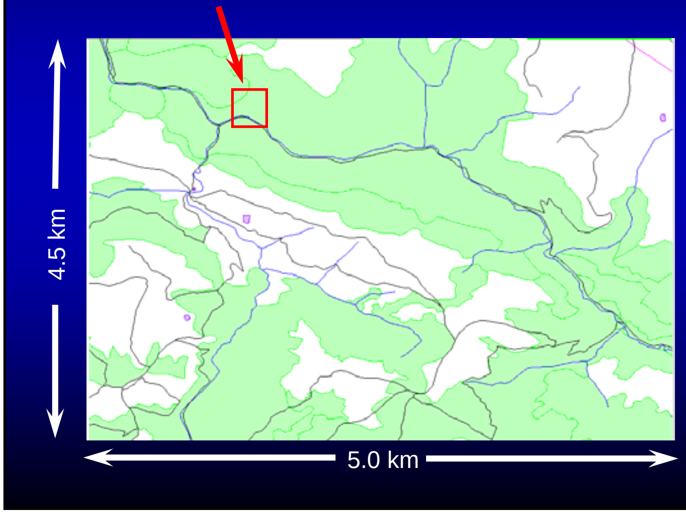
Topology

Topology requirements define the expectations for how geospatial features should be connected

- Breaks in road connectivity impact route planners
- Surface polygon adjacency failures impact mobility

Topology Along a Road Network

Area of Interest



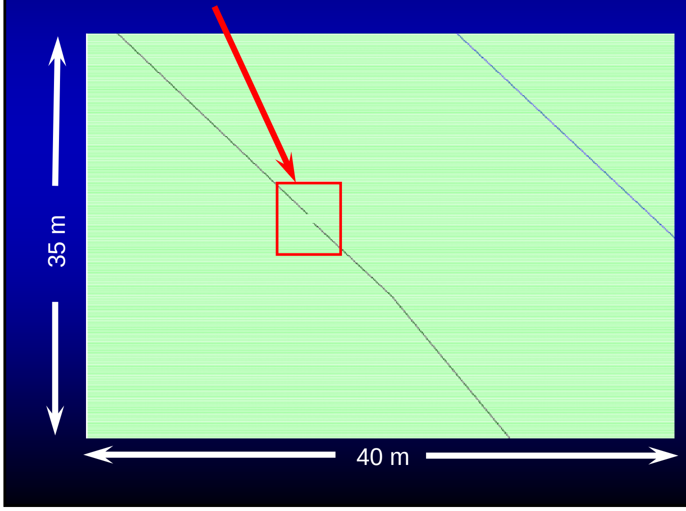
Roads – Black

Rivers – Blue

Vegetation – Green

Topology Along a Road Network

1 meter Gap between Road feature vertices



Roads – Black

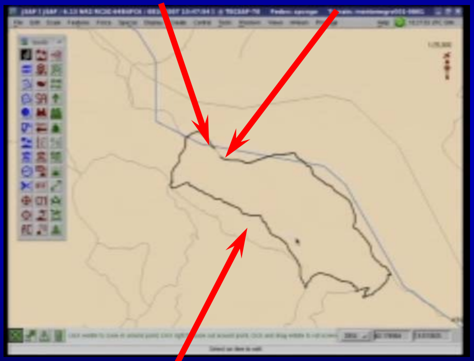
Rivers – Blue

Vegetation – Green

JSAF Routing Solution

JSAF on-road route

from here … to here …



JSAF takes a scenic tour
that avoids the gap

Joint Semi-Automated Forces (JSAF) routing solution for on-road travel (shown as dark black line)

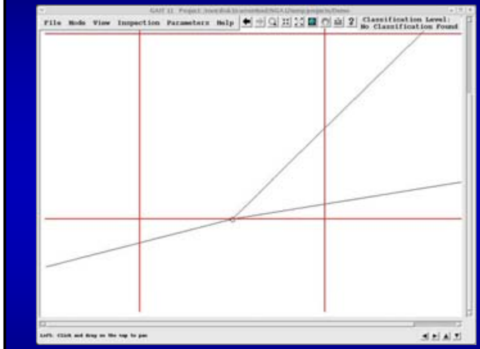
Situation Is Improving

- Multinational Geospatial Co-Production Program (MGCP)
  - MGCP includes 28 member nations committed to mapping much of the world landmass at scales of 1:50,000 or 1:100,000
  - Member nations contribute and withdraw data from the International Geospatial Warehouse
- Quality Assurance is an integral component of the MGCP program
  - “All Road Transportation Feature segments that are connected in reality shall be geometrically connected in the data.”
  - “Line features must not have kinks or kickbacks (collapsed loops)”

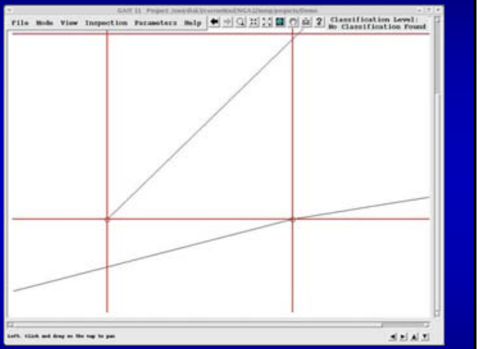
The Nature of Connections

- Interpreting the word “connected” – alternate views
  - Features appear connected in printed products or graphical display systems
  - Connected features share common vertex coordinates
    - Precision of representation
- Poor network connectivity is one example of the impact
- Coordinate transformation can amplify the associated problems
  - Network gaps are magnified
  - Features move

A Popular (but flawed) Remediation: Snap Coordinates to a Predefined Grid


File Mode View Inspection Parameters Help
Classification Level:
No Classification Found
Left: Click and drag on the map to pan

Decimeter Grid (in red) and two disconnected Road features (in black) before snapping


File Mode View Inspection Parameters Help
Classification Level:
No Classification Found
Left: Click and drag on the map to pan

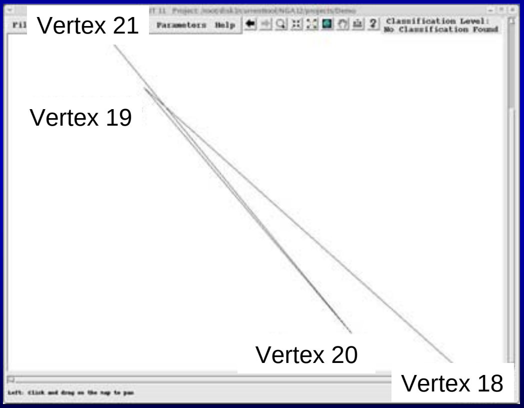
Snapping to the grid makes the gap much larger

Resolving the "Almost Equal" Coordinate Problem

- The popular strategy of snapping to a predefined grid is a flawed approach to correcting the problem
  - Can work in some cases
  - Only guarantee is minimization of maximum gap errors
    - Decreasing the grid spacing does not changes things, except to present more opportunities for snapping apart
- Only solution we have found is to snap 'nearby' coordinates together
  - The problem is defining 'nearby'

A Geometric Representation Problem: Line Kink

Line feature Kinks can result from inadvertently adding a point during manual data capture



Z-Shaped Kink in a Road feature

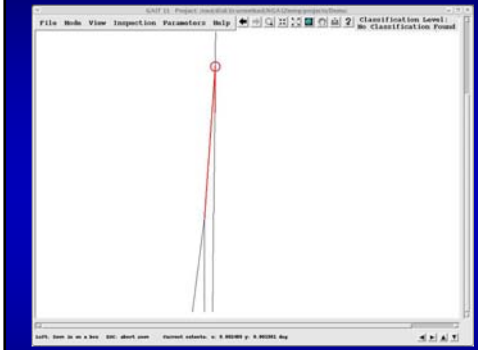
Distance from vertex 19 to vertex 20 in the Road feature is ~ 25 meters

Detecting Line Kink Errors

Thresholds	True Positives (TP)	False Positives (FP)	Threshold TP Rate	Cumulative TP Rate
0° < α ≤ 5°	701	236	74.8%	74.8%
5° < α ≤ 10°	676	1033	39.6%	52.0%
10° < α ≤ 15°	628	1929	24.6%	38.5%
15° < α ≤ 20°	744	3018	19.8%	30.7%

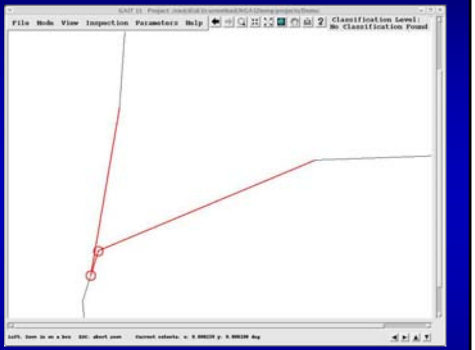
- Angle-only line Kink detector applied to 14 data sets
  - 1,399,785 line features
  - 51,625,949 line feature vertices
- Manual evaluation of the 8,965 reports of potential errors results in 2,749 TP and 6,216 FP – Not Good Enough!

These Geometric Problems Come in Multiple Varieties – Some are not Errors



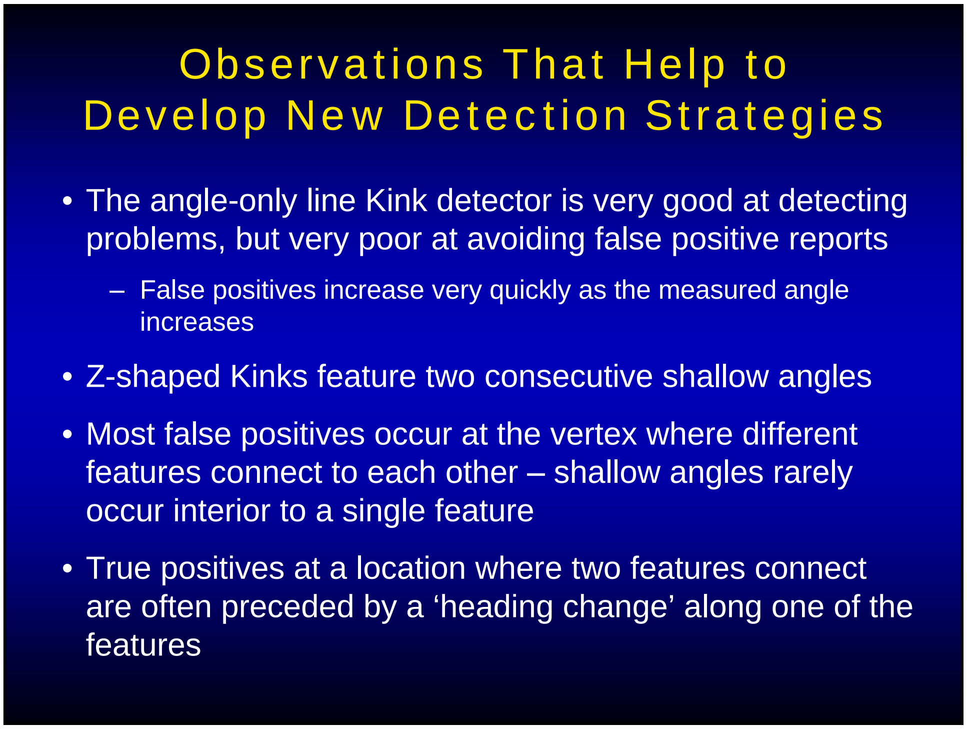
Shallow angle (Kink) where two Rail features connect

Error very unlikely



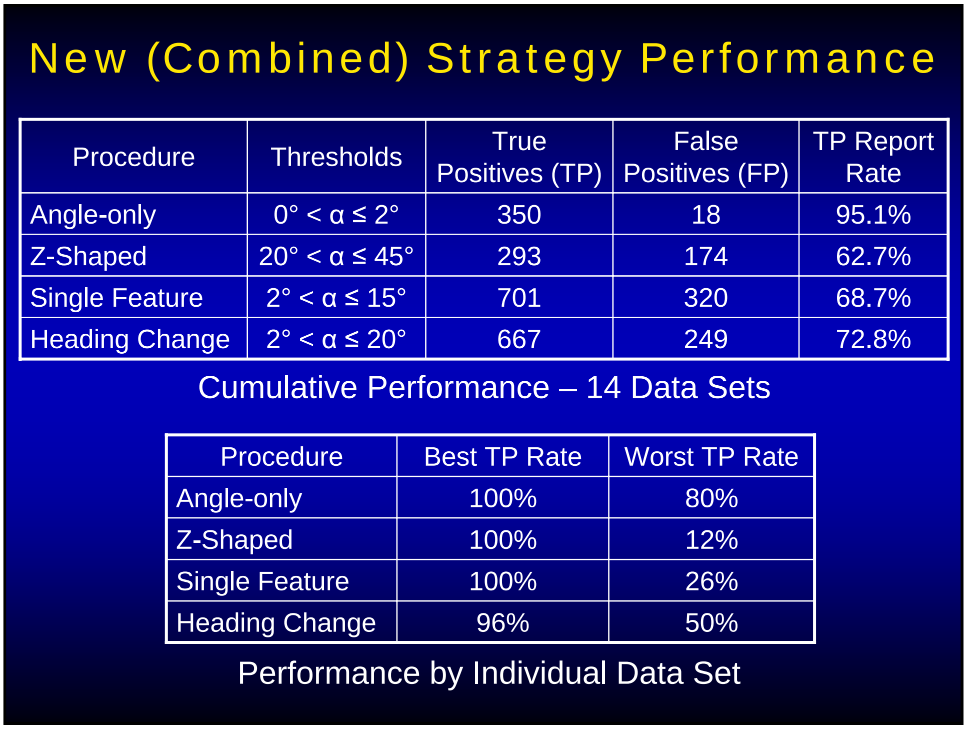
Shallow angle (Kink) where two Road features connect

Error highly probable



Observations That Help to Develop New Detection Strategies

- The angle-only line Kink detector is very good at detecting problems, but very poor at avoiding false positive reports
  - False positives increase very quickly as the measured angle increases
- Z-shaped Kinks feature two consecutive shallow angles
- Most false positives occur at the vertex where different features connect to each other – shallow angles rarely occur interior to a single feature
- True positives at a location where two features connect are often preceded by a ‘heading change’ along one of the features



New (Combined) Strategy Performance

Procedure	Thresholds	True Positives (TP)	False Positives (FP)	TP Report Rate
Angle-only	0° < α ≤ 2°	350	18	95.1%
Z-Shaped	20° < α ≤ 45°	293	174	62.7%
Single Feature	2° < α ≤ 15°	701	320	68.7%
Heading Change	2° < α ≤ 20°	667	249	72.8%

Cumulative Performance – 14 Data Sets

Procedure	Best TP Rate	Worst TP Rate
Angle-only	100%	80%
Z-Shaped	100%	12%
Single Feature	100%	26%
Heading Change	96%	50%

Performance by Individual Data Set

Summary

- The emergence of detailed quality specifications is unprecedented and will benefit all analytical users of geospatial data
- Imprecision in the specification is problematic
- The requirement to critically examine both specifications and data remains
- Structured experimentation can be applied to refine specifications

- There are solid grounds for optimism …